Lifebrain

D. 3.1. Development of a data storage and management system

Project title:	Healthy minds from 0-100 years: Optimising the use of European brain imaging cohorts
Due date of deliverable:	30th June, 2017
Submission date of deliverable:	28th June, 2017
Leader for this deliverable:	University of Oslo

This project has received funding from the *European Union's Horizon2020 research and innovation programme* under grant agreement *No 732592.*

Contributors to deliverable:	Name	Organisation	Role / Title
Deliverable Leader	Athanasia Monika Mowinckel	UiO	Researcher
Contributing Author(s)	Inge Amlien	UiO	Senior engineer
	Klaus Ebmeier	UOXF	WP3 leader
	Anders Fjell	UiO	WP6 leader
Reviewer(s)	William Baaré	RegionH	WP2 leader
	Sandra Düzel	MPIB	Researcher
	Simone Kühn	MPIB	Researcher
	Andreas Brandmeier	MPIB	Researcher
	Paolo Ghisletta	UNIGE	PI
	Sezen Cekic	UNIGE	Researcher
	Rik Henson	MRC	PI
	Rogier Kievit	MRC	Researcher
	Lorraine Tyler	UCAM	PI
	David Bartrés-Faz	UB	PI
	Christian Drevon	Vitas	WP5 leader
	Lars Nyberg	UmU	PI
	Mikael Stiernstedt	UmU	Administrative coordinator
Final review and approval	Kristine Walhovd	UiO	Coordinator, WP4 leader
	Barbara B. Friedman	UiO	Administrative coordinator

Document History				
Release	Date	Reason for Change	Status (Draft/In-review/Submitted	Distribution
1.0.	05.04.2017	First draft for the General Assembly	Sent for commenting	OneDrive
2.0.	05.15.2017	Revised version	Internal evaluation	OneDrive
3.0.	28.06.2017	Final version	Final revision	Participant portal

Dissemination level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of contents

Executive Summary

The aim of D.3.1. is to establish an international image sharing, storage and processing platform for management of neuroimaging and harmonized target variables. The proposed solution is to store data in a central location, at the University of Oslo with restricted access-permissions. Each data-contributor retains permissions to their own data, and no other partner has default access to other partners' data. This solution will complement meta-analytic strategies in Lifebrain.

The Services for Sensitive Data at UiO is user friendly and can be accessed by two-step verifications remotely through browsers or virtual machines. The service is also coupled with a parallel-process computing cluster, and has a multitude of software packages for analysis. This service provides the necessary functionality required for the consortium to start data-sharing promptly. The expenses for this service will be covered from the Lifebrain project budget of UiO.

List of acronyms / abbreviations

Information technology	IT
Extensible Neuroimaging Archive Toolkit	XNAT
Lifebrain Consortium Data Administrator	LCDA
Data manager	DM
General Assembly	GA
Material transfer agreement	MTA
Data transfer agreement	DTA
Services for sensitive data	TSD
University of Oslo	UiO
University Centre for Information Technology	USIT
Virtual Machine	VM
High performance computing	HPC
Hitachi Network Attached Storage	HNAS
Magnetic resonance imaging	MRI
Computed tomography	CT
Positron emission tomography	PET
Knowledge Management Committee	KMC
Secure File Transfer Protocol	SFTP
European Medical Information Framework	EMIF

Introduction

Description of deliverable

D3.1. Data storage and processing system to share data, manage data access and integrated analysis
Task 3.1: Development of a data storage and management system. Lead: UOXF; Participants: UiO, REGIONH, UmU, MPIB (M1-M6)

The task is to establish an international image sharing, storage and processing platform for management of neuroimaging and harmonized target variables. The platform will be provided by means of the open source Extensible Neuroimaging Archive Toolkit (XNAT). The infrastructure will include individual XNAT nodes at each contributing institution to securely host that site's data. Each XNAT instance will follow the harmonization protocol (WP2.1) making it easy for partners to search for and identify data for agreed collaboration projects via a central hub XNAT instance. The hub will allow to send query to each node and return summary information of qualifying data for the available cohorts within each node. Only summary information (i.e. no individual subject data) will be returned to the user, which then can be used to initiate data sharing and harmonised analysis.

Objectives

Develop a data management system that allows efficient storage, identification and sharing of sensitive project data across sites in accordance with strict data security legislations, necessary for the successful completion of WP2 and WP4.

Collaboration among partners

After the Lifebrain kick-off meeting in Brussels (16-18th January 2017) many discussions took place in the Lifebrain consortium, with the involvement of external experts on finding the optimal technical solution. Experiences were exchanged with a representative from the European Medical Information Framework (EMIF), an Innovative Medicines Initiative (IMI) programme that is coming into its` 5th year regarding data sharing (See minutes from 17.04.2017 in Annex 1). An assessment had been conducted by UOXF as well, among all participating sites on system needs, permission processes, ethical concerns and technical expectations toward the Lifebrain platform (See Annex 2). The conclusion was that the data should be stored in a central location with restricted access-permissions. The TSD system operated at the University of Oslo proved to be fitting all needs.

Therefore, the leader of the deliverable became UiO, working in close cooperation with UOXF, the original leader of task 3.1. The first draft was distributed to the General Assembly members 05.04.2017 for comments, which then were accommodated in the revised version.

1. Summary of the proposed solution

The Lifebrain-consortium needs a data-storage and analysis system where data is not only stored securely, but also is accessible for the consortium partners. In the initial Lifebrain application there was a storage solution where data was stored locally at each contributing site connected through XNAT hubs at each location. These hubs would then be accessible to Lifebrain partners, who could query the database and receive summaries of qualifying data from available cohorts at each site. While this solution would be preferential, it is not technically achievable to have up-and-running within the project timespan. It would furthermore require dedicated IT-services at each hub to develop the system. A central storage solution is, currently, the most feasible option. This will add to meta-analytic strategies also described in the GA, where each partner provides only effect size information, not requiring a central storage solution.

The proposed solution is to store data in a central location, at the University of Oslo, with restricted access-permissions. Each data-contributor retains permissions to their own data, and no other partner has default access to other partners' data. This solution requires two types of administrators: The Lifebrain Consortium Data Administrator and the Data Manager. The LCDA is the link between the consortium and the IT-service department at the University of Oslo. This person oversees relaying information about users, permissions, group-memberships etc. to the University IT-services. The DM (with selected persons as back-ups) has access to all data, and oversees the general data-handling. When the GA agrees to the start of a new project, the DM will extract and merge requested Lifebrain data-variables after ensuring that all necessary material and data transfer agreements are in place. Extracted data will be made available in project-specific folders, that can only be accessed by Lifebrain scientists involved in the specific sub-project. The DM is responsible for correcting reported data errors, for logging the changes that have been made, and providing affected projects with the corrected data. Therefore, none other than the DM has write-access to Site specific data bases/folders. A ticket-system to report errors to the DM will be developed, to ensure efficient management of such errors.

The University of Oslo has IT-services for the storage and analysis of sensitive data, see http://www.uio.no/english/services/it/research/storage/sensitive-data/index.html. This service is user friendly and can be accessed by two-step verifications remotely through browsers or virtual machines. The service is also coupled with a parallel-process computing cluster (Slurm), and has a multitude of software packages for analysis. This service provides the necessary functionality required for the consortium to start data-sharing promptly.

The UiO site will cover all Lifebrain's expenses for this service, which is estimated at a yearly 10 600 EUR as an operation cost.

2. Infrastructure

The TSD system has been developed at the University Centre for IT (USIT), University of Oslo (2011-14) and first launched on May 2014. The system has been set up to comply with the Norwegian national legislations concerning research on sensitive data, and most the hosted research projects contain health information that is directly or indirectly identifiable. These regulations follow European data protection regulations.

The basic layout of the system is a secure centralised service where data and backup (snapshots/mirrors and tape) are stored. Further, this central service provides a High-Performance Computing resource within the secure environment. Every project has its own set of Windows and Linux virtual machines within the central TSD system.

Authorized users can access the system from any location with internet access using a modern web browser.

The TSD infrastructure is outlined below, and interested parties may read the full whitepaper at: https://www.uio.no/english/services/it/research/sensitive-data/about/whitepaper_jan-2017.pdf.

2.1. Security

- All access demands two-step authentication (OAuth).
- Any changes in access rights needs approval from the LCDA.
- Dedicated storage, encrypted backups, and encrypted communication is used.
- Encryption keys are generated with a unique set of keys for each project/environment. These are stored on paper in a fire-proof safe in two separate locations.
- Data transfer in and out of the system is done via a special purpose file staging service.

2.2. Data Import/Export

Data transfer to and from the services is handled by a special purpose file staging service and the TSD project administrator controls access rights for all members of the Lifebrain TSD-project. By default, Lifebrain consortium TSD project members can transfer data to site import folders and within project-specific folders they may have access to. Only the Lifebrain consortium data administrator, data manager and other selected and approved members of the pxxx-export-group (see below) can transfer data out of the system. Copy/paste is disabled from the system, but users are still able to paste text into the system.

2.3. Backup

In addition to standard (encrypted) tape backup, snapshots of files and folders for the last three days are also created. This allows both long-term secure data storage, in addition to fast recovery of files by the users stored in snapshots, without administrator intervention.

2.4. Virtual Machines

Lifebrain consortium members will have the option to use both Windows (Server 2012 R2) and Linux (RHEL 6.x) virtual workstations. Resources (CPU/RAM) will be allocated to the VMs as needed. The virtual workstations will function as platforms for data analysis and submit hosts for the HPC cluster. The workstations have common productivity and scientific software installed by default or available, such as Microsoft Office, R, Matlab, SPSS, FSL, FreeSurfer. Other software packages can be installed manually (accessible only to user, or in a central location accessible to others) or with support from administrators.

Heavy analyses that require much computing resources (SEM, imaging processing etc.) should be run on the parallel processing system on TSD that may be accessed through the VMs (see below), not on the VMs themselves.

2.5. High Performance Computing

TSD has a dedicated HPC resource (Colossus), a compute cluster consisting of 72 nodes equipped with 20 cores and 64 GB RAM each in addition to 2 large memory nodes with 32 cores and 1 TB RAM each. In total 1500+ CPU cores. The HPC is set up similar as the larger Abel computing facility at UiO (http://www.uio.no/english/services/it/research/hpc/abel/more/index.html), with shared HNAS storage disks and Fraunhofer BeeGFS for work partitions, connected by infiniband, and running the SLurm queue system. The HPC nodes have a range of scientific software and libraries installed by default (http://www.uio.no/english/services/it/research/hpc/abel/help/software/).

2.6. Imaging Database

The TSD infrastructure provides support for databases (e.g. PostgreSQL), and we have the option to deploy an XNAT server if needed. XNAT is an open-source platform designed to facilitate common management and productivity tasks for imaging and associated data. It consists of a repository for raw and post-processed images, a database to store metadata and non-imaging measures. XNAT supports all common imaging methods (e.g. MRI, CT, PET), and can be extended to support related data such as demographics, behaviour data, genetics, and offers granular access control mechanisms. The server will be located inside the secure TSD environment, accessible from the virtual machines after login with two-factor authentication only. UiO will provide basic back-end support for running and maintaining an XNAT instance. Development of functionality will rely on expertise from partners or external personnel.

3. Procedures

3.1. Lifebrain consortium data administrator

This role refers to the Lifebrain consortium at the TSD system level. There are several separate projects being operated on TSD infrastructure, and the LCDA is responsible for administrative tasks pertaining to the operation of the Lifebrain consortium inside TSD. This includes resolving issues of group memberships, approving registration of users, requesting resources, software etc., which will dynamically change throughout the project period.

3.2. Gaining access

New users apply for access to the system through a designated person at his or hers partner site, who after confirmation from the data protection officer provides the users with an electronic registration form that is then forwarded to the LCDA. TSD executes the routines needed for user creation. Login information consisting of a username in the form of p274-xxxxx, a password, QR code for one-time-code initialization or Yubikey for users lacking smartphone, is sent by post to the intended user, or placed in the home folder of the LCDA who then forwards the information in a secure fashion to the intended user.

3.3. Export privileges

Only users who are members of the export-file-group can export files. Request for export permission can be granted by the GA. The LCDA informs the TSD administrator who adds the user to pxxx-export-group. The LCDA and data protection officer keep an updated list of users with export rights. Data are imported and exported through a special file-staging area, using sftp.

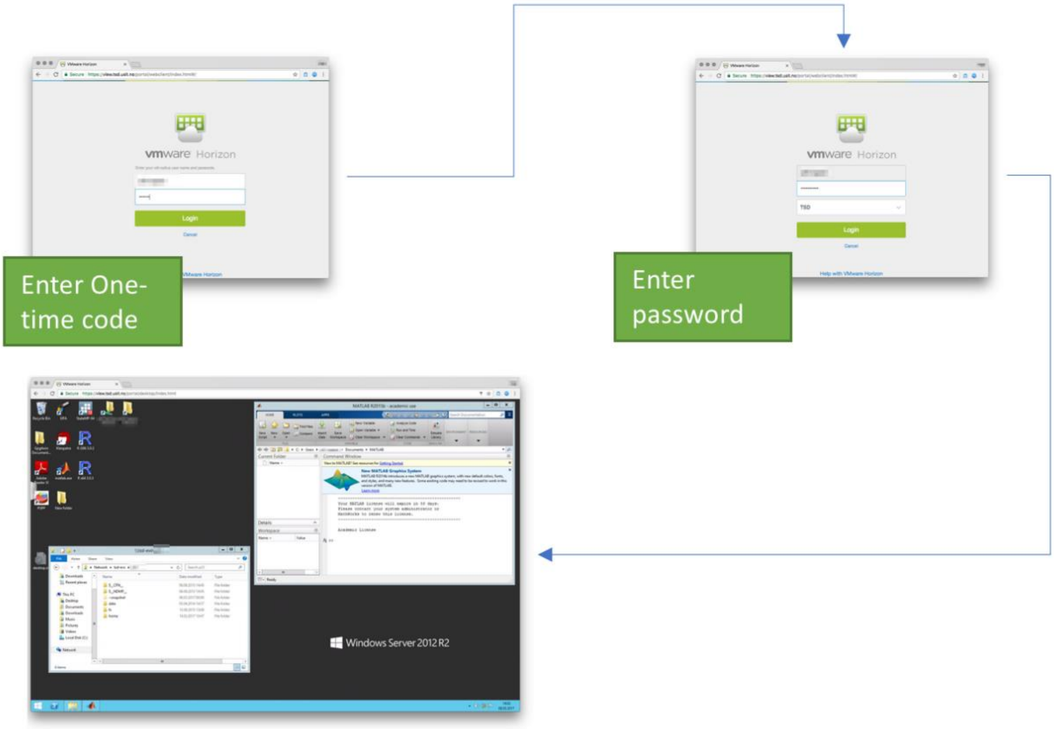
Details of the procedures are found at https://www.uio.no/english/services/it/research/storage/sensitive-data/use-tsd/import-export/index.html.

3.4. Logging in to the system

To log in to the system, the user must have an active p274-xxxxx account, a password, and a means to generate one-time codes for two-factor authentication (smartphone or Yubikey). The user may choose to log in to either a Windows or a Linux VM through a simple login process.

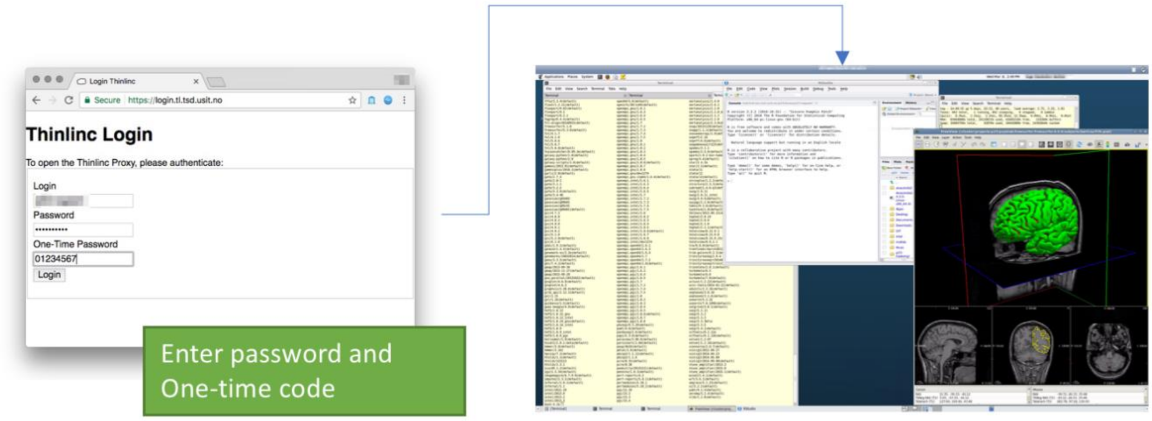
Windows VM: Open a web browser (Internet Explorer, Firefox, Chrome) and connect to https://view.tsd.usit.no

Figure 1. Login in Windows VM



Linux VM: Open a web browser (Internet Explorer, Firefox, Chrome) and connect to https://login.tl.tsd.usit.no

Figure 2. Login in Linux VM



Thinlinc Login

To open the Thinlinc Proxy, please authenticate:

Login [redacted]

Password ••••••••••

One-Time Password 01234567

Enter password and One-time code

4. Data organization

4.1. Folder structure

Folders will be organized in a pre-determined way, that ensures secure data management. The project area (p274) will include 3 folders:

- **Sites** - Data organized by site and study/time point.
- **Projects** – Folders for specific Lifebrain sub-projects
- **Logs** – Error logs, data dictionaries, summaries and descriptives of available data.

4.2. Data manager

The DM has master access to all files. The DM populates site specific data bases/folders with data provided by individual partner sites and corrects identified data errors and logs the changes made. Moreover, the DM extracts and merges Lifebrain data for GA approved Lifebrain sub-projects, after notification by the Knowledge Management Committee (KMC) and making sure that necessary MTAs and DTAs are present.

4.3. Folder access

Access to folders is organized by means of user-groups. Personnel from each Site are members of Site user groups, and will have read access to their own Site's data, but not to that of other Sites. Site members will also have read and write access to Site import folders, which is the staging area for incoming data that the DM quality controls and adds to the database. Lifebrain consortium user groups are determined and created at the initiation of each new project, such that only persons collaborating on specific projects will have access to data and analyses of the specific project. All users will have read access to the Log folder, to enquire about the existence of data for a project idea or to check if data errors have been detected and rectified.

Figure 3. Folder access

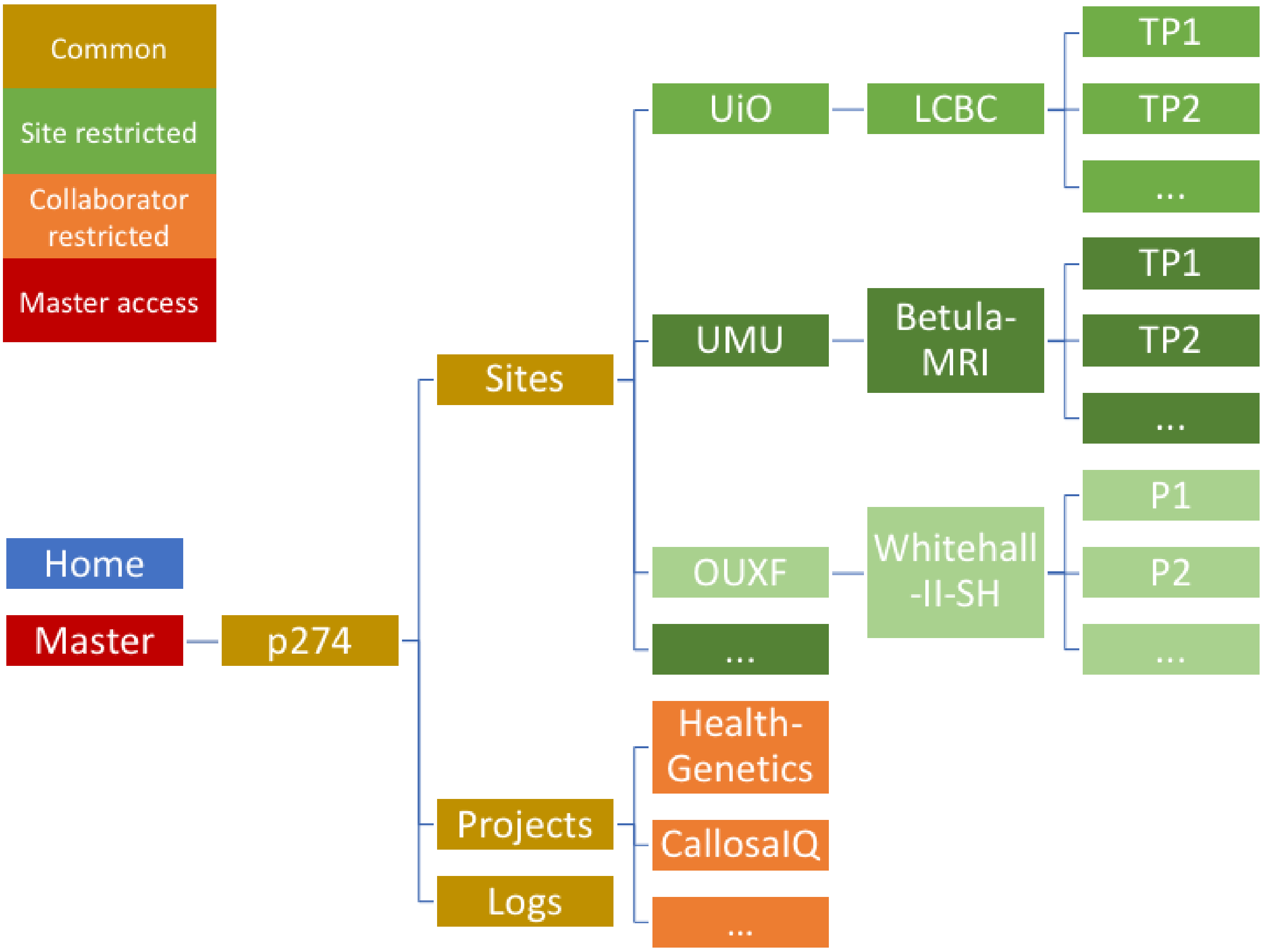


Figure 3 Folder structure. Home folder: Every user with access to the system has a private home-directory, i.e. only the user can read (r), write (w) and execute (x) on that folder.

Green area: Access restricted site specific data. Site specific personnel have read, but not write, access permission.

Orange area: When a new project has been approved by the GA, a new project folder will be created, and the specified data will be added to the directory by the Data Manager (DM). These folders are only accessible (read, write and execute) by Lifebrain scientists listed as project members.

Yellow area: common area, accessible to all Lifebrain scientists.

Master access (Red): The DM can access all site data. The DM is the only person who can write data into site specific databases/folders (green) and uniformly correct reported data-errors and log the changes that are made to the data. The DM will, after notification by the KMC, create project-level (orange) folders for Lifebrain consortiums that have been approved by the GA, and add the required Lifebrain data to the project folder after first ensuring that all the necessary MTAs and DTA's are in place.

4.4. Importing data and error correction

Sites will be required to organize their data files in the manner directed by WP2.1 (data harmonization). UiO will assist with conversion and data-check scripts. New data files will be uploaded to site specific import-folders which is a part of the two-stage process of securely entering data to TSD. Site personnel have write access to this folder. After notification, the DM does quality checks and moves the new data into the site-specific folders. Each site retains full read access to their own data. Write access in the Site data-folders is restricted to the DM. Moreover, the DM ensures that any corrections to data are controlled and logged, and that all affected projects are informed and provided with corrected data. As solutions for bona-fide database solutions (e.g. PostgreSQL) are developed, there will be a transition from using spreadsheets to full database solutions.

4.5. Initiating projects

Project folders are created upon the approval from the GA to commence a new project and extract data. The KMC notifies the DM regarding the new project and the Lifebrain scientists that need access to the project folder, in addition to necessary information to identify the requested Lifebrain data. The DM ensures that all MTAs and DTAs with respect to the requested project data are present before extracting and making data available to project members in a newly created project folder. Numeric data will be made available in open formats (csv, data, etc.), which are easily read by a multitude of different software's.

4.6. Data handling

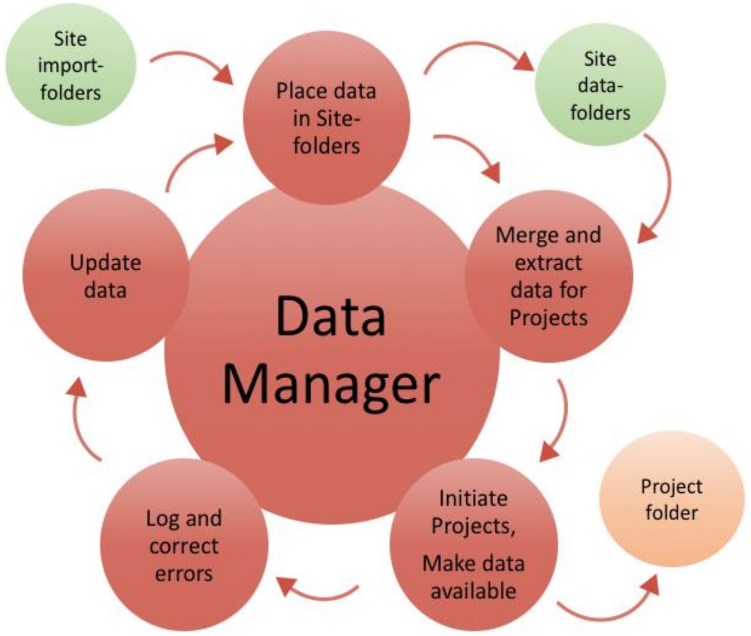
The complex nature of consortia like the current, requires a balance between the creation of optimal and practical solutions for data sharing. There is a need to first build a system that enables the partners to begin analysing data, while a more fine-tuned system for sharing data across sites is developed.

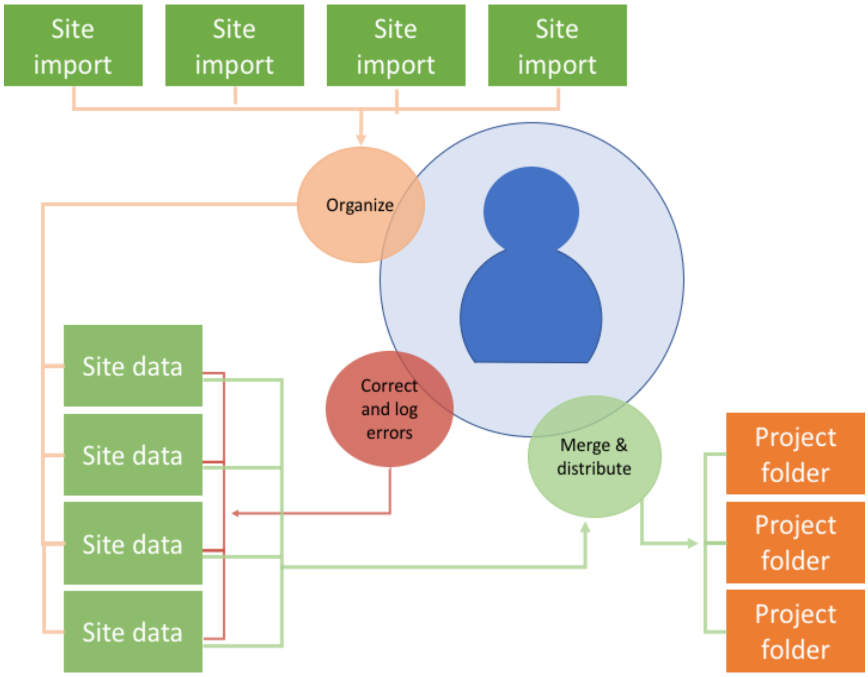
As a pragmatic solution, data handling is here proposed to undergo several stages, incrementally increasing data complexity and handling-flexibility.

1) Numerical data in large spreadsheets handled by the DM
2) Imaging data in XNAT (PostgreSQL interface)
3) Integration of XNAT and spreadsheets
4) Continued development of database solutions (e.g. PostgreSQL)

To make certain data available for consortium partners as soon as possible, a preliminary instance of working with numerical data in large spreadsheets will be created. This would be handled though R statistical software and git version control for tracking script versions. Given proper data-harmonization, data extraction should be easily handled by the DM, by the import and export of data in data tables. The initial focus of working with spreadsheets will make it possible to start working with the data, while the handling of the data using a data base software solution will be developed. Initial spreadsheets may also include summary data from already processed MRI-data. The second step would be to initiate the imaging database in XNAT. The complexity of imaging data requires more intricate handling, which is assumed to need more time to make available than simpler numerical data. The scope of stage 3 is to create a full integration of the larger spreadsheets and the XNAT sever. At this stage is has not been decided if this will be a pure XNAT solution or a combination of XNAT (specifically developed for handling imaging data) and other database software that has been specifically developed for handling non-imaging data (e.g. other PostgreSQL interfaces). Stage 4 is an ongoing process throughout the consortium to maintain and develop database solutions.

Figure 4. Data handling process


Site import-folders
Place data in Site-folders
Site data-folders
Update data
Data Manager
Merge and extract data for Projects
Project folder
Log and correct errors
Initiate Projects,
Make data available


Site import
Site import
Site import
Site import
Organize
Site data
Site data
Site data
Site data
Correct and log errors
Merge & distribute
Project folder
Project folder
Project folder

Annex

Annex 1 Minutes of meeting with EMIF

- Skype, 17.03.2017

Stephanie - Postdoc. - EMIF

Attendees: Kristine Walhovd, Anders Fjell, Klaus Ebmeier, Athanasia Monika Mowinckel

Notes from talk with Stephanie

Harmonization - Dichotomize, z-scores.

- They also have the raw data

What platform do they use?

- TranSmart, either to store the data, and secure them and you can do some analyses and export data. - https://wiki.transmartfoundation.org/
  - They put their data centrally
  - Data is sent to a central place, where harmonization is done
  - Then sent to a company that uploads it to tranSmart, and does some extra checks on compatibility
- Still do some manual transformations with statistical programs

Someone with experience with the data needs to harmonize and merge, as it has to do with understanding what can and should be harmonized and merged together. IT-specialists without prior knowledge of the content of the data will find it difficult to merge in consistent ways.

- After skype WP3 meeting , 03.17.2017

Attendees: Kristine Walhovd, Anders Fjell, Klaus Ebmeier, Athanasia Monika Mowinckel

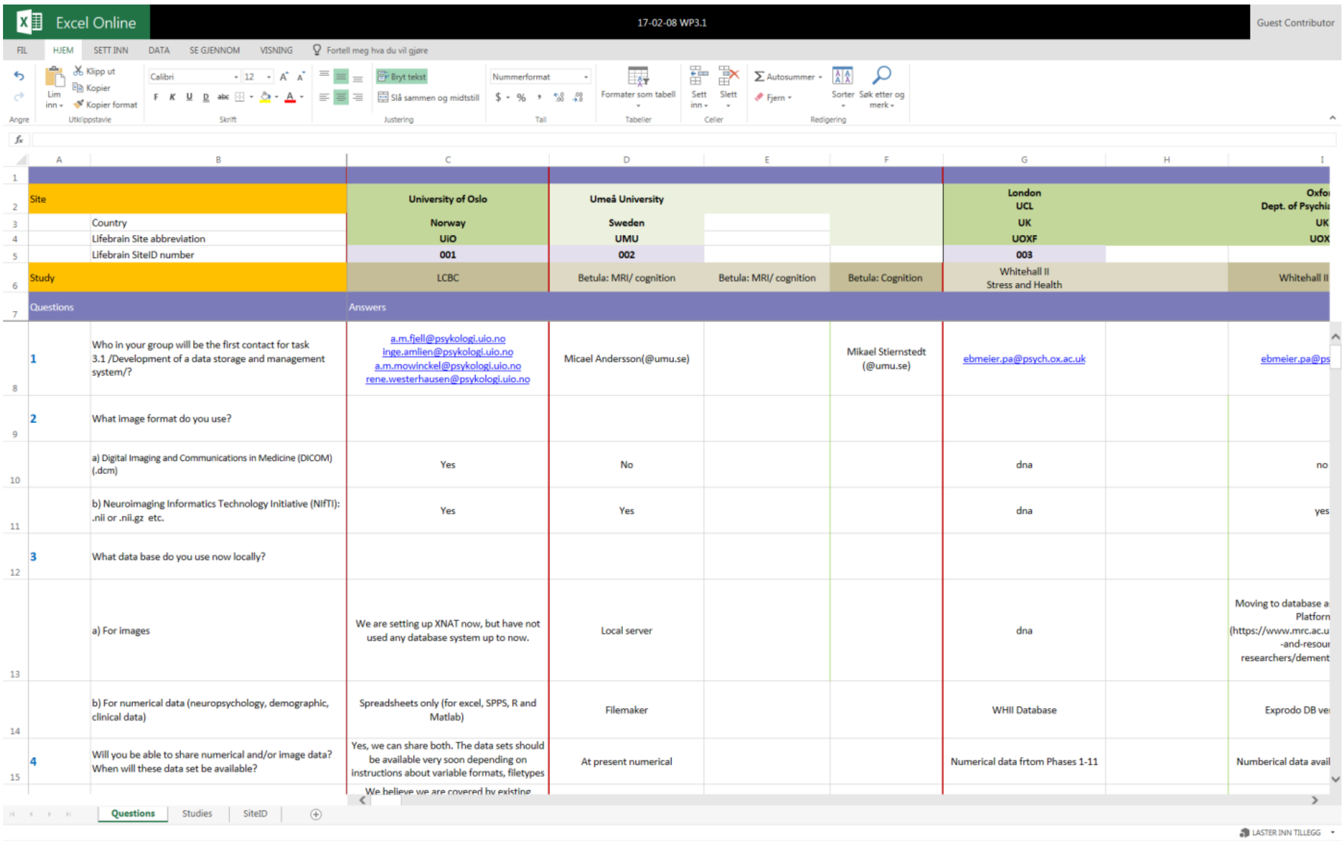
Notes from after-talk with Lifebrain attendees:

Do we want to go directly for a database structure right away, or do we want to just start with the "simpler" solution of working with spreadsheets?

- In order to get data available and for researchers to have some starting point of data to work with, it is best to start with the simple solution, and work top-down.
- If we get simple data-spreadsheets going, that can be merged and data extracted using scripts, the merged data may easily be adaptable to "proper" databases at a later stage.
- LCBC has the competency to get a preliminary set-up running, adapting work that has been employed at their Center.

The WP3.1 proposal set forward by LCBC is being revised after comments from partners, and will be made available for further input and the GA promptly.

Annex 2: Assessment on technical parameters of local/sites databases



	A	B	C	D	E	F	G	H	I
1									
2	Site		**University of Oslo**	**Umeå University**			**London** **UCL**		**Oxfo** **Dept. of Psychi**
3		Country	**Norway**	**Sweden**			**UK**		**UK**
4		Lifebrain Site abbreviation	**UiO**	**UMU**			**UOXF**		**UOX**
5		Lifebrain SiteID number	**001**	**002**			**003**		
6	Study		LCBC	Betula: MRI/ cognition	Betula: MRI/ cognition	Betula: Cognition	Whitehall II Stress and Health		Whitehall II
7	Questions		Answers						
8	1	Who in your group will be the first contact for task 3.1 /Development of a data storage and management system/?	a.m.fjell@psykologi.uio.no inge.amlien@psykologi.uio.no a.m.mowinckel@psykologi.uio.no rene.westerhausen@psykologi.uio.no	Micael Andersson(@umu.se)		Mikael Stiernstedt (@umu.se)	ebmeier.pa@psych.ox.ac.uk		ebmeier.pa@ps
9	2	What image format do you use?							
10		a) Digital Imaging and Communications in Medicine (DICOM) (.dcm)	Yes	No			dna		no
11		b) Neuroimaging Informatics Technology Initiative (NIfTI): .nii or .nii.gz etc.	Yes	Yes			dna		yes
12	3	What data base do you use now locally?							
13		a) For images	We are setting up XNAT now, but have not used any database system up to now.	Local server			dna		Moving to database a Platforn (https://www.mrc.ac.u -and-resour researchers/dement
14		b) For numerical data (neuropsychology, demographic, clinical data)	Spreadsheets only (for excel, SPPS, R and Matlab)	Filemaker			WHII Database		Exprodo DB ve
15	4	Will you be able to share numerical and/or image data? When will these data set be available?	Yes, we can share both. The data sets should be available very soon depending on instructions about variable formats, filetypes	At present numerical			Numerical data frtom Phases 1-11		Numberical data avail
			We believe we are covered by existing						

Annex 3: Limitations of person identifiable data-information

To ensure the data cannot identify the participating subjects, certain restrictions on the data will be enforced. These restrictions include the omission of data regarding (the list is not exhaustive):

- Name
- Address
- E-mail address
- Telephone numbers
- Official ID-numbers (national insurance, passports etc.)
- IP addresses
- Face, fingerprint or handwriting
- Credit card numbers
- Digital identity
- Date of birth (to be replaced with month and year only)
- Birthplace (to be replaced with general location)
- Genetic information
- Login names, screen names, nicknames and handles
- Non-de-faced MRIs
- PID (e.g. DNA), at least until ethical approvals are in place for such storage

Annex 4: Available Linux software for super-computing

The computing cluster, Colossus, has several working versions of software installed simultaneously, and users may load the version they prefer working on. USIT has user guides in loading specific software versions: http://www.uio.no/english/services/it/research/hpc/abel/help/user-guide/modules.html. Users are encouraged to specify which software version to use, and to install the (unlicensed) software they wish to use. USIT also maintains a cran-repository within TSD, that is periodically updated (http://www.uio.no/english/services/it/research/sensitive-data/use-tsd/software/r/index.html). It is also possible for users to download and import R-libraries into TSD, if they require the newest (or developer) versions of the packages. This requires no extra permissions. Users may choose to store library packages in different (personal or shared) folders, depending on which R-version and package-versions they prefer to use.

This following is a list of software available on the compute cluster and the Linux VMs through the #module load command. The Linux and MS Windows VMs have additional softwares available, such as Office, SPSS, zip, text editors, etc. Updated list including softwares available on Windows and Linux VMs is to be found at https://www.uio.no/english/services/it/research/storage/sensitive-data/use-tsd/software/index.html.

Programming languages

GHC
GPU/NVIDIA/CUDA
Julia
MATLAB
OCaml
Perl modules
Python 2
Python 3
R

Compilers
Intel
Open64
Portland

Debuggers
Scalasca
TotalView

Libraries
Boost
FFTW
GSL
HDF5
ICU
IntelMPI
netCDF
OpenMPI

Visualisation
Gnuplot
Graphviz

General software
bzip2
CFITSIO
CMake
curl and libcurl
hpczip
ImageMagick
pbzip2
pcre
Perf-reports
Spark
SPRNG
Subversion (svn)
taskfarm
tre
wget
xz

Statistics
R
Stata

Biology/ bioinformatics/ Imaging
454apps
alleleCount
ABySS
AmpliconNoise
BEAGLE
BEAST
BLAST
BLAST+
bowtie2
cdbfasta
CD-HIT
CEGMA
Clearcut
ClonalOrigin
ClustalW
denoiser
dosageconverter
FastTree
flash
fqgrep
Freesurfer
FSL
GARLI
GATK
Geneid
hisat2
HMMER
HUMAnN2
IMPUTE2
Infernal
Interproscan
LAMARC
MAFFT
MACS2
MAGMA
mcmcphase
metaxa2
MGLTools
Microbiome Utilities
Migrate
MIRA
Molden
mothur
MrBayes
MUSCLE
NCL
ngsplot
Novoalign
Orthograph
PAML
Pandaseq
ParsInsert
PAUP
PennSNV
PhyloBayes
PhyML
Picard-tools
PLINK
PLINKSEQ
pplacer
ProtTest
QIIME
RAxML
rtax
SHAPEIT
SPAdes
structure
STAR
Stringtie
Subread
swarm
TransDecoder
TREEFINDER
Trinotate
UCLUST
UNPHASED
USEARCH
Velvet
vsearch
Wise2 (formerly GeneWise)

Chemistry
ADF
AMBER
AutoDock
AutoDock Vina
CP2K
Gaussian
LAMMPS
MaterialsStudio
NAMD
QuantumEspresso
VASP

Computational linguistics
VISL CG-3
Geo Sciences
ESyS-Particle
FLEXPART
OpenIFS
WRF & WRF CHE